Approches basées sur les réseaux Bayésiens pour la prédiction d'attaques sévères

Karim Tabia, Philippe Leray

Approches basées sur les réseaux Bayésiens pour la prédiction d'attaques sévères

Karim Tabia* — **Philippe Leray***

* *Equipe COnnaissances et Décision (COD) - LINA UMR CNRS 6241*
Ecole Polytechnique de l'Université de Nantes
{Karim.Tabia,philippe.Leray}@univ-nantes.fr

RÉSUMÉ. *Dans cet article, nous proposons des modélisations basées sur les réseaux Bayésiens de type multi-nets pour un problème majeur en détection d'intrusions et corrélation d'alertes. Dans un premier temps, nous proposons un modèle prédictif permettant de prédire de futures attaques sévères en corrélant les alertes générées par un ensemble systèmes de détéction d'intrusions (IDSs). L'avantage de ce modèle réside dans l'utilisation d'un multi-net pour modéliser les corrélations propres à chaques catégories d'activités malveillantes à anticiper. Dans un deuxième temps, nous proposons de tenir compte de la fiabilté des IDSs en utilisant la méthode Virtual Evidence de Pearl. Dans un troisième temps, pour contrôler les taux de prédiction/taux de fausses alertes sous jacent, nous proposons d'implémenter le principe de prédiction avec rejet permettant à l'administrateur de fixer à l'avance le taux de fausses alertes à ne par dépasser. Enfin, nous rapportant les résultats expérimentaux validant nos modélisations réalisés sur des logs d'alertes IDMEF réelles collectées dans le cadre du projet PLACID.*

ABSTRACT. *In this paper, we propose a Bayesian multi-net approach allowing to i) better model the various malicious activities to predict, ii) handle the reliability of multiple intrusion detection systems (IDSs) when predicting severe attacks and iii) provide a flexible and efficient approach especially designed to limit the false alarm rates by controlling the confidence of the prediction model. Our experimental studies carried out on a real and representative IDMEF alert corpus collected in the framework of PLACID project show very interesting performances regarding the tradeoffs between the prediction rates and the corresponding false alarm ones.*

MOTS-CLÉS : *Réseaux Bayésiens, multi-nets, classification avec rejet, corrélation d'alertes*

KEYWORDS: *Bayesian networks, multi-nets, classification with reject option, alert correlation*

1. Introduction

En détection d'intrusions, on déploie souvent plusieurs solutions et produits de sécurité afin d'accroître les taux de détection et de couverture en exploitant les complémentarités mutuelles des solutions adoptées. Cependant, les systèmes de détection d'intrusion (IDSs par suite) sont bien connus pour générer de grandes quantités d'alertes dont la plupart sont fausses et redondantes. Ce problème est dû à plusieurs raisons telles que des paramètres inappropriés pour les IDSs, etc. (Tjhai *et al.*, 2008). Afin de faire face à de telles quantités d'alertes, des approches de corrélation d'alerte ont été proposées (Debar *et al.*, 2001).
La corrélation d'alertes est l'analyse des alertes déclenchées par un ou plusieurs IDSs afin de fournir une vue *synthétique* et de emphhaut niveau des événements malveillants *intéressants* ciblant le système d'information. Ainsi, les approches de corrélation d'alertes visent soit à réduire le nombre d'alertes pour en éliminer les redondantes (Debar *et al.*, 2001) ou la détection de plans attaques (Ning *et al.*, 2002) où les différentes alertes correspondent à l'exécution d'un plan d'attaque s'étalant sur plusieurs étapes. Récemment, les auteurs dans (Benferhat *et al.*, 2008b) proposent de sélectionner parmi les alertes générées uniquement celles qui correspondent aux connaissances et préférences de l'administrateur du système.
Dans cet article, nous proposons une approche visant d'abord à réduire le nombre d'alertes générées par les IDSs puis prédire efficacement les attaques sévères. Les principaux avantages de cette approche sont : i) Minimum d'intervention de l'expert , ii) Anticipation des attaques sévères, ce qui permet de prendre les contre-mesures appropriées avant de subir des dégâts et iii) Traitement efficace de la fiabilité des IDSs et un contrôle du taux de prédiction d'attaques sévères/taux de fausses alertes. Notons que la plupart des approches de corrélation d'alerte existantes ne tiennent pas compte de la fiabilité des IDSs. Dans notre approche, la prédiction d'attaques sévères est considérée comme un problème de classification. Pour modéliser et exploiter la fiabilité des IDSs, nous utilisons la méthode dite "Virtual Evidence" de Pearl (Pearl, 1988) qui permet de raisonner en présence d'observations incertaines dans le cadre des réseaux Bayésiens. Afin de contrôler les taux de prédiction/taux de fausses alertes, notre approche permet de rejeter les séquences d'alertes lorsque la confiance de notre modèle de corrélation d'alerte n'est pas suffisante pour faire une bonne prédiction. Comme nous le constaterons dans nos études expérimentales, notre approche permet de réduire considérablement le taux de fausses alertes en garantissant des taux de prédictions d'attaques sévères très intéressants.

2. Corrélation d'alertes : bref aperçu

La corrélation d'alertes (Debar *et al.*, 2001) consiste à analyser les alertes générées par un ou plusieurs IDSs et éventuellement par d'autres outils de sécurité afin de fournir une vue *synthétique* et de *haut niveau* des événements malveillants détéctés. Les données en entrée pour les outils de corrélation d'alertes sont collectées auprès de diverses sources telles que les IDSs, les pare-feux, logs de serveurs web, etc. La corrélation des alertes signalées par de multiples analyseurs présente plusieurs avantages

notamment l'augmentation de la couverture et l'exploitation des complémentarités mutuelles des outils utilisés. Dans notre cas, une alerte est un message textuel généré par un IDS quand une attaque ou une activité anormale est détectée. Il contient une identification/nom de l'activité détectée, sa catégorie, un niveau de sévérité (traduisant le niveau de gravité ou de danger encouru), l'adresse IP de l'attaquant, l'adresse IP de la victime, etc. on peut résumer les principaux objectifs de la corrélation d'alertes dans les points suivants :

1) *Réduction d'alerte et d'élimination des alertes redondante :* L'objectif ici est d'éliminer la redondance des alertes par agrégation ou fusion d'alertes similaires (Debar *et al.*, 2001). En effet, les IDSs génèrent souvent de grandes quantités d'alertes redondantes en raison de la multiplicité des IDSs et la répétitivité de certains événements malveillants tels que les scans, les attaques DoS, DDoS, etc.

2) *Détection de plans d'attaques :* La plupart des IDSs ne rapportent que des événements élémentaires alors que plusieurs attaques malveillantes s'étalent sur plusieurs étapes où chacune peut être révélée par une alerte. Détecter les plans attaques requiert l'analyse des relations entre plusieurs alertes.

3) *Filtrage et priorisation d'alertes :* Parmi les énormes quantités d'alertes générées, des administrateurs doivent sélectionner un sous-ensemble d'alertes en fonction de leur dangerosité et du contexte de chaque système d'information. Le filtrage et priorisation d'alertes ont pour but de présenter aux administrateurs seulement les alertes qu'ils souhaitent analyser en priorité (Benferhat *et al.*, 2008b).

Dans la littérature, les approches de corrélation d'alerte sont souvent regroupées en approches fondées basée sur la similarité entre alertes (Debar *et al.*, 2001), sur le scénarios d'attaque prédéfinis (Ning *et al.*, 2002) et approches statistiques (Valdes *et al.*, 2001). Dans cet article, nous nous intéressons à la prédiction d'attaques sévères (qui représent un grave danger si jamais elles sont lancées et exécutées avec succès) qui peut être considérée comme une variante de détection de plans d'attaque.

2.1. *Fiabilité des IDSs : Une question cruciale*

Le problème le plus important recontré par les utilisateurs d'IDSs concerne le nombre exhorbitant de fausses alertes[1]. En effet, tous les IDSs sont connus pour leurs taux élevés de fausses alarmes. Dans une évaluation expérimentale de Snort (Tjhai *et al.*, 2008), les auteurs ont conclu que 96% des alertes générées sont fausses. Par conséquent, il est clair que la prise en compte de la fiabilité des IDSs est une question importante pour des tâches de corrélation d'alertes, comme la prédiction d'attaques sévères, l'objet de ce travail. Par exemple, si l'on sait que 90% des alertes déclenchées par un IDS concernant une attaque donnée sont fausses, cette information devrait alors être prise en compte si ces alertes doivent être exploitées par l'outil de corrélation d'alertes. Notons qu'il n y a, à notre connaissance, aucun travail sur la modélisation et la prise en compte de la fiabilité des IDSs pour la prédiction d'attaques sévères.

1. Une fausse alerte correspond à activité légitime signalée par erreur comme action malveillante par l'IDS

3. Multi-nets pour la prédiction d'attaques sévères

3.1. *Réseaux Bayésiens pour la classification*

Les réseaux bayésiens (également appelés modèles graphiques probabilistes, systèmes experts probabilistes, etc.) sont des outils compacts, expressifs et très puissants pour l'extraction, la modélisation et le raisonnement avec des informations incertaines et complexes (Jensen *et al.*, 2007). Un réseau Bayésien est spécifié par i) *Une composante graphique* consistant en un DAG (Directed Acyclic Graph) qui permet de représenter les variables du problème et leurs relations (dépendances) et ii) *Une composante probabiliste* consistant en un ensemble de table de probabilités conditionnelles permettant de quantifier l'incertitude relative aux relations de dépendance entre les variables. Les réseaux Bayésiens sont utilisés pour différents types d'inférence tels que le maximum à a posteriori (MAP), l'explication la plus probable (MPE), etc. Quant aux applications, ils sont utilisés comme outils d'extraction de connaissances à partir de données, modèles prédiction comme la classification, diagnostic, etc.
La classification est une tâche très utilisée dès lors qu'il faut répartir des objets, individus, etc. caractérisés chaun par un ensemble d'attributs en classes prédéfinies. Elle consiste à prédire la valeur d'une variable cible C (non observée) sur la base des valeurs des variables observées, appelées attributs. Ainsi, étant donné les variables observées A_1,..,A_n décrivant les objets à classer, il est demandé de prédire la *vraie* valeur de la variable de classe C. La classification avec un réseau Bayésien est un type particulier d'inférence MAP consistant à calculer l'instance de la variable de classe ayant la plus grande probabilité a posteriori étant donné l'instance du vecteur attribut $a_1a_2..a_n$. La règle de classification basée sur le maximum a posteriori dans les réseaux Bayésiens peut être écrite de la manière suivante :

$$Classe = argmax_{c_k \in C}(p(c_i/a_1a_2..a_n)), \quad [1]$$

où le terme $p(c_i/a_1a_2..a_n)$ représente la probabilité a posteriori d'avoir c_i étant l'instance $a_1a_2..a_n$. Cette probabilité est calculée avec la règle de Bayes comme suit :

$$p(c_i/a_1a_2..a_n) = \frac{p(a_1a_2..a_n/c_i) * p(c_i)}{p(a_1a_2..a_n)} \quad [2]$$

Il est à préciser que la plupart des travaux utilisant les réseaux Bayésiens pour la classification utilisent des modèles naïfs ou semi-naïf (comme les réseaux TAN ou BAN) (Cheng *et al.*, 2001). Ces modèles font des hypothèses fortes pour simplifier l'apprentissage du modèle à partir de données alors que les classifieurs Bayésiens non naïfs ont besoin d'apprentissage de la structure et des paramètres (constructions des tables de probabilités conditionnelles).
En sécurité informatique, les réseaux Bayésiens sont très utilisés et dans de nombreux domaines. Plus particulièrement, les classifieurs Bayésiens sont utilisés dans la détection d'intrusions (Valdes *et al.*, 2000). En corrélation d'alertes, une approche basée sur un réseau Bayésien est utilisée dans (Valdes *et al.*, 2001) pour la fusion d'alertes. Les classifieurs Bayésiens sont également utilisés dans (Benferhat *et al.*, 2008a) (Faour *et al.*, 2006) où les auteurs utilisent principalement des modèles naïfs et TAN pour la détection des plans d'attaques. Notons que tous les travaux sur la détection de plans

d'attaques utilisent des modèles prédictifs naïfs semi-naïfs et qu'il n'existe, à notre connaissance, aucun travail traitant de la prise en compte de la fiabilité les IDSs en corrélation d'alrtes. Dans ce qui suit, nous proposons un modèle prédictif basé sur un multi-net pour la prédiction d'attaques sévères.

3.2. *Multi-nets pour la prédiction d'attaques sévères*

Un classieur Bayésien standard est composé d'un réseau unique qui encode les relations de dépendance entre les variables. Ces dépendances sont évaluées sur toutes les données d'apprentissage sans distinction de classe. Cependant, force est de constater que ces relations ne sont pas les mêmes dans les différentes classes. Plus particulièrement, dans notre application (prédiction d'attaques sévères), chaque attaque sévère est statistiquement corrélée uniquement avec un petit ensemble spécifique d'autres alertes (la plupart du temps parce que plusieurs attaques sont menées par des vers et des scripts qui entrainent les mêmes événements malveillants). Comme nous le verrons dans nos études expérimentales, la modélisation de corrélations locales permet à une meilleure estimation des probabilités a posteriori pour la classification. Notons aussi que la plupart des algorithmes d'apprentissage de structure utilisent des tests et mesures statistiques pour identifier les corrélations fortes. Par conséquent, quand les classes sont déséquilibrées, le réseau obtenu encode principalement les corrélations de la classe majoritaire. Dans un multi-net, chaque instance de classe c_i est représentée par un réseau N_{ci} encodant seulement les relations de dépendance locales estimées sur les instances d'apprentissage appartenant exclusivement à la classe c_i. Les fréquences des différentes classes peuvent être codées par un nœud racine C, associé à la variable de classe ou tout simplement par une distribution de probabilité locales comme dans (Cheng *et al.*, 2001).
La prédiction d'attaques sévères consiste à analyser des séquences d'alertes ou d'événements d'audit afin de prédire de futures attaques sévères. Dans ce travail, la prédiction d'attaques sévères est modélisée comme un problème de classification où les variables sont définies comme suit :

1) **Prédicteurs (attributs) :** L'ensemble des prédicteurs (variables observées) est composé des alertes *pertinentes* pour la prédiction des attaques sévères. Ainsi, à chaque alerte pertinente $Alert_i$, on associe une variable A_i signalant le fait que l'alerte $Alert_i$ a été générée/absente dans la séquence d'alertes à analyser.

2) **Variable à prédire (classe) :** La variable de classe C représente les attaques sévères à prédire. Son domaine comprend toutes les attaques sévères $Attack_1$,..,$Attack_n$ à prédire en plus d'une autre instance RAS (rien à signaler) représentant les séquences d'alertes qui ne sont pas suivies par des attaques sévères.

Les principaux avantages de cette approche sont :

1) *Minimum d'intervention experte :* l'expert n'a qu'à identifier les attaques sévères qu'il veut prédire. Pour déterminer les alertes pertinentes à utiliser, il peut recourir auxs méthodes de sélection de variables. En outre, les modèles prédictifs sont automatiquement construits à partir de l'historique des alertes collectées.

2) *Facilité de déploiement :* il suffit de prétraiter les alertes générées en temps réel pour en extraire les attributs décrivant chaque séquence d'alerts à analyser et d'effectuer en temps réel de prédiction d'attaques sévères.

3) *Facilité de mise-à-jour :* le modèle prédictif peut facilement être mis à jour par le ré-apprentissage du modèle sur des données plus appropriées.

Dans la section suivante, nous proposons une approche pour tenir compte de la fiabilité des IDSs lors de la prédiction des attaques sévères.

4. Exploitation de la fiabilité des IDSs

Le problème de la prise en compte de la fiabilité des IDSs est crucial et peut être vu dans le cadre des réseaux Bayésiens comme un problème d'inférence en présence d'observations incertaines. Pearl (Pearl, 1988) a proposé une méthode très simple pour représenter et raisonner avec des informations incertaines. Présentons d'abord cette méthode dite "Virtual Evidence" et montrons comment peut-on l'utiliser pour la fiabilité des IDSs.

4.1. *Raisonnement en présence d'observations incertaines*

La méthode Virtual Evidence de Pearl (Pearl, 1988) offre un intuitif et naturel pour le raisonnement avec des observations incertaines. Dans cette méthode, l'incertitude indique la confiance que l'on a sur le fait que les valeurs observées sont les vraies valeurs. Dans notre contexte, si un IDS déclenche une alerte et nous savons (à partir de l'expérience passée, par exemple) que cet événement est une fausse alerte dans 95% des cas, alors nous sommes en présence d'observation incertaine. L'idée principale de la méthode de Pearl est de reporter l'incertitude relative à l'observation incertaine E sur un évènement virtuel R de telle sorte que l'incertitude pesant sur E est spécifiée comme la vraisemblance de R dans le contexte de E. A titre d'exemple, si un test médical est fiable à 80% et que si un test donné est positif, cette incertitude sera représentée par une variable T (pour test) dans le contexte de la vraie variable M (pour maladie) qui ne peut pas être observée directement. L'incertitude du test médical sera alors représentée par la vraisemblance du test dans le contexte de M, c'est à dire par une table de probabilités conditionnelles $P(T|M)$, ce qui se traduit dans le réseau par l'ajout d'un nœeud fils pour chaque variable suscepble d'être observée avec incertitude. Désormais, la variable observée est T et on peut inférer M en tenant compre de l'incertitude des tests.

4.2. *Prédiction d'attaques sévères avec prise en compte de la fiabilité des IDSs*

Afin d'appliquer la méthode de Pearl pour tenir compte de la fiabilité des IDSs, nous devons d'abord évaluer la fiabilité de ces IDSs par le biais d'évaluations empiriques en examinant pour chaque type d'alerte déclenchée par un IDS, la proportion de

vraies/fausses alertes[2]. Ainsi, après avoir évalué la fiabilité des IDSs quant aux alertes A_1,..,A_n, le traitement de l'incertitude pesant sur une séquence d'alerte, on procède comme suit :

1) Pour chaque variable A_i associée à une alertes utilisée comme prédicteur, nous ajoutons une nœud fils R_i pour coder l'incertitude sur A_i. Le domaine de R_i est D_{R_i}=$\{0, 1\}$ où la valeur 0 est utilisée pour représenter l'incertitude concernant le cas A_i=0 (l'alerte A_i n'a pas été rapportée) tandis que la valeur 1 est utilisée pour coder l'incertitude dans le cas où A_i=1 (l'alerte A_i a été déclenchée).

2) Chaque distribution de probabilités conditionnelles $p(R_i/A_i)$ encode la fiabilité que les valeurs observées soient effectivement de vraies attaques. Par exemple, la probabilité $p(R_i$=$1/A_i$=$1)$ code la probabilité que l'observation R_i=1 est effectivement due à une attaque réelle.

Lors de l'analyse d'une séquence d'alerte $r_1..r_n$ (une instance des variables R_1,..,R_n), nous calculons $argmax_{ci}(p(c_k/r_1..r_n))$ pour prédire la classe de la séquence d'alertes analysée. Il est à souligner qu'en pratique, il est moins compliqué d'évaluer le taux de vrais/faux taux positifs que d'évaluer les taux de vrais/faux négatifs car pour ces derniers, il faut analyser l'ensemble des activités (par exemple, tout le trafic réseau) afin d'évaluer la proportion d'attaques qui n'ont pas été détectée par les IDSs. Dans la section suivante, nous proposons une solution pour contrôler les taux de prédiction/ taux de fausses alertes.

5. Option de rejet pour contrôler les taux de prédiction/taux de fausses alertes

5.1. *Classification avec rejet*

La classification avec option de rejet consiste à rejeter (ne pas classifier) les objets qui sont susceptibles d'être mal classifiés. A titre d'exemple, si la probabilité que l'instance à classifier est en deça d'un certain seuil (fixé par l'expert), l'instance en question sera rejetée pour éviter qu'elle ne soit mal classifiée. Dans la litérature, il existe deux type de rejet, qui sont implémentée de différentes manières :

1) **Rejet de distance :** Cette situation apparaît lorsque l'objet à classifier n'appartient à aucune des classes modélisées par le classifieur. Cela peut être dû à l'existence d'une classe non représentée ou que l'objet en question est $outlier$[3]. Ici, le rejet de distance sert à délimiter les classes apprises/modélisées par le système de classification et permet, ainsi, de rejeter ce qui est en dehors de ses compétences. Dans la pratique, cette solution est implémentée en mesurant le degré d'appartenance ou distance de l'objet à classifier aux différentes classes. Un seuil est fixé en deça duquel l'objet à classifier est rejeté.

2) **Rejet d'ambiguïté (ou de confusion) :** Dans ce cas, l'objet à classifier $appartient$ à plusieurs classes en même temps, ce qui rend le classifieur confus. Cela

2. Un expert peut aussi subjectivement (par expérience) fixer la fiabilité des IDSs qui composent son infrastructure de détection d'intrusions

3. Ce cas peut se présenter par exemple lorsqu'un capteur ou outil de mesure fournit des données erronées et abérrantes

peut être causé par le fait que les classes apprises ne sont pas disjointes. Cette situation peut être constaté par exemple par des exemples d'apprentissage ayant une même description (même valeurs d'attributs) mais appartenant tantôt à une classe tantôt à une autre. Ce type de rejet est implémenté en détectant les instances qui sont proches de plusieurs classes en même temps. Plusieurs techniques sont utilisées pour mettre en oeuvre ce rejet. En effet, dans (Chow, 1970), l'auteur a proposé d'utiliser des degrés de probabilité a posteriori de l'instance à classifier dans les différentes classes.

5.2. *Contrôle des taux de prédiction/taux de fausses alertes*

Les classieurs Bayésiens sont naturellement adaptés pour implémenter la classification avec rejet puisque que la classification se fait par le calcul des probabilités a posteriori d'instances de classe étant donnés l'objet à classer. Chaque probabilité $p(c_i/a_1..a_n)$ peut être interprétée comme de la confiance du classifieur que l'instance à classer $a_1..a_n$ appartient à la classe c_i. Dans notre application, nous sommes intéressés par le contrôle du taux de prédiction d'attaques sévères/taux de fausses alertes selon les besoins de chaque utilisateur final. Par exemple, un utilisateur peut vouloir un outil de corrélation d'alertes avec une confiance élevée (pour garantir un minimum de fausses alertes). Cet objectif requiert le rejet des séquences d'alertes pour lesquelles l'outil n'est pas très confiant. Définissons la confiance φ dans notre cas comme la valeur (non signée) de l'écart de la probabilité que l'instance à classer $a_1a_2..a_n$ ne sera pas suivie par une attaque sévère et la plus grande probabilité qu'elle sera suivie par l'une des attaques sévères modélisées. Cette mesure est proposée dans (Leray *et al.*, 2000) pour évaluer la confiance d'un classieur Bayésien.

$$\varphi(a_1..a_n) = |p(c_i = RAS/a_1..a_n) - \max_{c_i \neq RAS}(p(c_i/a_1..a_n)| \quad [3]$$

Dans la formule de l'Equation 3, c_i=0 désigne l'instance de classe qui représente les séquences d'alerte qui ne sont pas suivies par des attaques sévères et toute instance de classe $c_i \neq 0$ désigne une attaque sévère à prédire. La valeur de $\varphi(a_1a_2..a_n)$ donne une estimation de la confiance du classifieur que la séquence d'alertes sera/ne sera pas suivie d'une attaque sévère. Ainsi, un utilisateur qui veut rejeter toutes les séquences d'alerte lorsque la probabilité d'être suivie par une attaque sévère n'est pas deux fois plus grande que la probabilité de ne pas l'être se fera en fixant le seuil de rejet L à la valeur 1/3. Notons que les probabilités a posteriori des classes étant donnée la séquence d'alertes à analyser doit être normalisée afin d'être utilisés avec un tel seuil. Désormais, la règle de décision Bayésienne de l'équation 1 sera reformulée comme suit :

$$Classe = \begin{cases} argmax_{c_k \in D_C}(p(c_i/a_1..a_n)) & if \varphi(a_1..a_n) > L \\ \emptyset & otherwise \end{cases} \quad [4]$$

La valeur $\emptyset$ représente la décision de rejet, c'est-à-dire "l'instance à classer est rejetée parce que la condition $\varphi(a_1..a_n) > L$ n'est pas satisfaite. Dans la section suivante, nous fournissons nos résultats expérimentaux sur les approches proposées dans cet article pour la détection d'attaques sévères.

6. Etudes expérimentales

Nos études expérimentales sont menées sur des logs d'alertes réelles générés par l'IDS Snort déployé sur un réseau de campus universitaires. Ces logs d'alertes représentent trois mois d'activité collectés au cours de l'été 2007 dans le cadre du projet PLACID [4]. Les données en entrée pour notre système consiste en des alertes générées par Snort au format IDMEF. Afin de formater ces alertes en données CSV qui peuvent être utilisées pour la construction de nos modèles, nous avons développé un outil de pré-traitement d'alertes. Ainsi, chaque séquence d'alerte CSV résume les alertes IDMEF survenues pendant une fenêtre horaire dont la largeur est fixée par l'utilisateur.

6.1. *Données d'apprentissage et de test*

Nos jeux de données sont obtenus à partir des logs d'alertes IDMEF réelles. Nous avons d'abord formater le premier mois de d'alertes collectées afin de construire le jeu d'apprentissage et le deuxième mois pour construire le jeu de test. La Table 1 fournit des détails sur les attaques sévères utilisées dans nos expérimentations. Parmi les attaques sévères détectées par Snort, nous avons sélectionné 9 attaques Web sévères à prédire sur la base des alertes de moindres sévérités qui précèdent ou préparent ces attaques Web sévères à prédire. Toutes ces attaques sont associés à un niveau de sévérité élevé et ciblent soit des serveurs Web soit des applications Web associées. En cas de réussite, de telles attaques conduisent à l'exécution de codes arbitraires et le contrôle total du système attaqué.

		Jeu d'apprentissage		Jeu de test	
Sid	Nom Snort de l'alerte	#	%	#	%
1091	WEB-MISC ICQ Webfront HTTP DOS	87	0,18%	6	0,01%
2002	WEB-PHP remote include path	50	0,10%	231	0,47%
2229	WEB-PHP viewtopic.php access	5169	10,42%	1580	3,20%
1012	WEB-IIS fpcount attempt	3	0,01%	10	0,02%
1256	WEB-IIS CodeRed v2 root.exe access	2	0,004%	3	0,01%
1497	WEB-MISC cross site scripting attempt	5602	11,30%	7347	14,90%
2436	WEB-CLIENT Microsoft wmf metafile access	145	0,29%	53	0,11%
1831	WEB-MISC jigsaw dos attempt	659	1,33%	153	0,31%
1054	WEB-MISC weblogic/tomcat .jsp view source...	3412	6,88 %	3885	7,88%

Tableau 1. *Distributions des attaques sévères dans les jeux d'apprentissage et de test*

6.2. *Expérimentation 1 : Multi-net pour la détection d'attaques sévères*

Afin d'évaluer l'efficacité de notre multi-net pour la prédiction d'attaques sévères, nous le comparons avec un arbre de décision C4.5 (Quinlan, 1993), un classifieur Bayésien naïf et un autre classifieur Bayésien construit en utilisant l'algorithme MWST (Chow *et al.*, 1968) permettant de construire des structures sous forme d'arbres (Francois *et al.*, 2004). Pour le multi-net, nous avons également utilisé l'algorithme MWST pour construire les différents réseaux représentant chaque attaque

4. `http:\placid.insa-rouen.fr`

sévère à détecter. Enfin, nous donnons les résultats de VE-multi-net, le multi-net implémentant la méthode de Pearl pour tenir compte de la fiabilité de Snort. Les ré-

Sid	Nom Snort de l'alerte	C4.5	Bayes naïf	MWST	Multi-net	VE-multinet
1091	WEB-MISC ICQ Webfront HTTP DOS	0%	0%	0%	0%	0%
2002	WEB-PHP remote include path	26,41%	26,41%	26,84%	26,84%	25,97%
2229	WEB-PHP viewtopic.php access	72,97%	74,81%	74,94%	72,15%	74,30%
1012	WEB-IIS fpcount attempt	0%	10%	0%	0%	0%
1256	WEB-IIS CodeRed v2 root.exe access	0%	0%	0%	0%	0%
1497	WEB-MISC cross site scripting attempt	92,02%	95,62%	95,62%	95,92%	93,32%
2436	WEB-CLIENT Microsoft wmf metafile..	56,60%	60,38%	60,38%	56,60%	56,60%
1831	WEB-MISC jigsaw dos attempt	50,3%	54,90%	54,25%	46,41%	37,25%
1054	WEB-MISC weblogic/tomcat .jsp view..	38,74%	41,36%	38,69%	47,77%	41,83%
	Taux de prédiction	72,26%	75,32%	74,53%	**76,92%**	73,88%
	Taux de fausses alertes	1,66%	3,21%	3,10%	1,58%	**0,74%**

Tableau 2. *Résultats de C4.5, Bayes naïf, MWST, multi-net et VE-multi-net*

sultats du Tableau 2 montrent que les performances de notre multi-net surpasse à la fois l'arbre de décision C4.5, Bayes naïf et le classifieur MWST eu égard aux taux de prédictions et les taux de fausses alertes sous-jacents. En particulier, le multi-net prédit 76,92% des attaques sévères à un taux de fausses alertes de 1,58% (29 fausses alertes/jour) tandis que Bayes naïf (resp. MWST) ont déclenché 3.21 % (58 fausses alertes/jour) (resp. 3,10 % (56 fausses alertes/jour)). Cette performance est due à la meilleure modélisation de chaque attaque sévère menant à une meilleure estimation de la vraisemblance des séquences d'alerte à analyser. Quant au classifieur *VE-multi-net*, il permet d'obtenir des taux de prédiction comparables à ceux du multi-net mais réduit considérablement le taux de fausses alertes jusqu'à **0,74%** (le taux de fausses alertes est passé de 29 à 13 fausses alertes/jour). Précisons que ce résultat est obtenu par la prise en compte de la fiabilité de trois alertes seulement (celles ayant sid=882, sid=1288 et sid=1852) et qui constituent la majorité de fausses alertes déclenchées par Snort dans nos jeux de données (voir (Tjhai *et al.*, 2008) pour une analyse de ces fausses alertes déclenchées par Snort). Ces résultats sont très prometteurs mais nécessitent une évaluation rigoureuse de la fiabilité de Snort. Il est important de noter que les quatre classieurs n'ont pas réussi à prédire certaines attaques sévères principalement en raison du déséquilibre de ces classes dans notre jeu d'apprentissage. Tel est par exemple la raison pour laquelle l'attaque sévère avec sid= 1256 qui n'est représentée dans le jeu d'apprentissage que par 2 instances.

6.3. *Expérimentation 2 : classification avec rejet pour contrôler les taux de prédiction/fausses alertes*

Dans l'Expérimentation 1, nous avons montré que la multi-net offre le meilleur compromis taux prédiction/taux de fausses alertes et permet naturellement de prendre en compte la fiabilité de Snort. Toutefois, dans les situations réelles, il est important de disposer de moyens pour la configuration du modèle prédictif utilisé de manière à ne pas dépasser un certain taux de fausses alertes. Dans l'Expérimentation 2, nous donnos nos résultats sur l'utilisation du rejet pour le contrôle du taux de prédiction/taux de fausses alertes. Ces résultats sont obtenus en définissant différents niveaux de confiance L et nous avons utilisé le même multi-net que dans l'Expérimentation 1.

La Figure 1 donne la courbe ROC de notre modèle évalué sur le jeu de test du Tableau 1. Le graphique de la Figure 1 montre la fluctuation du taux de prédiction du modèle

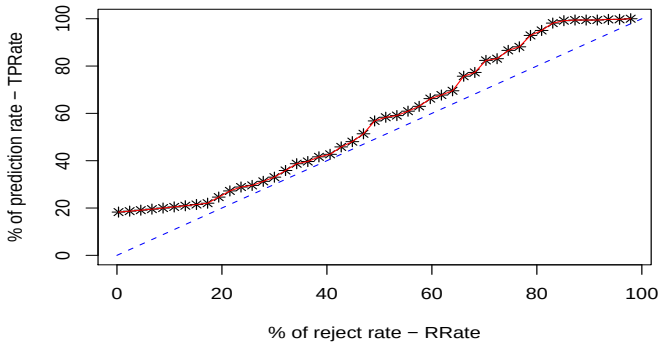

Figure 1. *Courbe ROC du VE-multi-net : taux de prédiction ($TPRate$) en fonction du taux de rejet ($RRate$)*

en fonction du taux de rejet. Cette courbe ROC montre que l'option de rejet améliore le pouvoir prédictif du modèle et permet à un expert de savoir quel taux de rejet correspond à chaque taux de prédiction qu'il peut vouloir garantir. Clairement, la Figure 1 montre que l'exploitation de l'option de rejet est efficace que ce même modèle n'utilisant pas cette option pour contrôler le taux de prédiction/taux de fausses alertes. Les résultats expérimentaux fournis dans cette section montrent clairement l'efficacité de notre modèle prédictif basé sur un multi-net pour la prédiction d'attaques sévères.

7. Conclusions

Le présent article a traité une question importante dans le domaine de la corrélation d'alertes consistant en la prédiction d'attaques sévères avec la prise en compte de la fiabilité des IDSs. Plus précisément, nous avons proposé un modèle prédictif basé sur un réseau Bayésien multi-net pour une meilleure modélistion des attaques sévères à prédire. Pour prendre en compte la fiabilité des IDSs, nous avons proposé l'implémentation de la méthode Virtual Evidence de Pearl qui permet de raisonner en présences d'observations incertaines. Afin de permettre le contrôle des taux de prédiction/taux de fausses alertes, nous avons proposé une approche basée sur l'option de rejet permettant de rejeter des séquences d'alertes lorsque le modèle de prédiction n'a pas assez de confiance pour faire de bonnes prédictions. Nos résultats expérimentaux sont très prometteurs, surtout lorsque l'on évalue de manière appropriée la fiabilité des IDSs et si l'on construit de bon jeux d'apprentissage. Comme continuations possibles de ce travail, il est très intéressant de prendre en compte la fiabilité des IDSs lors de la construction des jeux de données dans le cadre de l'apprentissage de réseaux Bayésiens à partir de données imparfaites.

8. Remerciements

Ce travail a été réalisé dans le cadre du projet ARN SETIN 2006 PLACID (`http://placid.insa-rouen.fr/`).

9. Bibliographie

Benferhat S., Kenaza T., Mokhtari A., « False alert filtering and detection of high severe alerts using Naive Bayes », *Computer Security Conference(CSC'08)*, South Carolina, apr, 2008a.

Benferhat S., Sedki K., « Alert Correlation based on a Logical Handling of Administrator Preferences and Knowledge », *International Conference on Security and Cryptography(SECRYPT'08)*, Porto, Portugal, p. 50-56, jul, 2008b.

Cheng J., Greiner R., « Learning Bayesian Belief Network Classifiers : Algorithms and System », *14th Conference of the Canadian Society on Computational Studies of Intelligence*, Springer-Verlag, London, UK, p. 141-151, 2001.

Chow C., « On optimum recognition error and reject tradeoff », *IEEE Transactions on Information Theory*, vol. 16, n° 1, p. 41-46, Jan, 1970.

Chow C., Liu C., « Approximating discrete probability distributions with dependence trees », *Information Theory, IEEE Transactions on*, vol. 14, n° 3, p. 462-467, 1968.

Debar H., Wespi A., « Aggregation and Correlation of Intrusion-Detection Alerts », *Recent Advances in Intrusion Detection*, Springer, London, UK, p. 85-103, 2001.

Faour A., Leray P., « A SOM and bayesian network architecture for alert filtering in network intrusion detection systems », *RTS - Conference on Real-Time and Embedded Systems*, p. 1161-1166, 2006.

Francois O., Leray P., « Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens », *Proceedings of 14eme Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, France, p. 1453-1460, 2004.

Jensen F. V., Nielsen T. D., *Bayesian Networks and Decision Graphs (Information Science and Statistics)*, Springer, June, 2007.

Leray P., Zaragoza H., d'Alch-Buc F., « Pertinence des Mesures de Confiance en Classification », *12eme Congres Francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Articifielle (RFIA 2000)*, Paris, France, p. 267-276, 2000.

Ning P., Cui Y., Reeves D. S., « Constructing attack scenarios through correlation of intrusion alerts », *9th ACM conference on Computer and communications security*, ACM, NY, USA, p. 245-254, 2002.

Pearl J., *Probabilistic reasoning in intelligent systems : networks of plausible inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

Quinlan J. R., *C4.5 : programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

Tjhai G. C., Papadaki M., Furnell S., Clarke N. L., « Investigating the problem of IDS false alarms : An experimental study using Snort », *23rd International Information Security Conference SEC 2008*, p. 253-267, 2008.

Valdes A., Skinner K., « Adaptive, Model-Based Monitoring for Cyber Attack Detection », *Recent Advances in Intrusion Detection*, p. 80-92, 2000.

Valdes A., Skinner K., « Probabilistic Alert Correlation », *Recent Advances in Intrusion Detection*, Springer-Verlag, London, UK, p. 54-68, 2001.